ELF and language change at the individual level

Vetchinnikova, Svetlana

Cambridge University Press
2020

Vetchinnikova , S & Hiltunen , T 2020 , ELF and language change at the individual level . in A Mauranen & S Vetchinnikova (eds) , Language Change : The impact of English as a lingua franca . Cambridge University Press , Cambridge , pp. 205-233 . https://doi.org/10.1017/9781108675000.012

http://hdl.handle.net/10138/323934
https://doi.org/10.1017/9781108675000.012

This is a preprint of Vetchinnikova, Svetlana & Turo Hiltunen. In press. ELF and language change at the individual level. In Anna Mauranen & Svetlana Vetchinnikova (eds.), *Language Change: The impact of English as a Lingua Franca*. Cambridge: Cambridge University Press.

ELF and language change at the individual level

Svetlana Vetchinnikova and Turo Hiltunen, University of Helsinki

Abstract

In this chapter we attempt to separate the communal and the individual level of language representation and explore how linguistic regularities emerge at each of them. We sample one communal and ten individual corpora of language use from the same ELF environment and examine to what extent syntactic structure, priming and chunking influence linguistic choice in each corpus by looking at the variation between contracted and full forms (*it is/it's*). We find clear differences in how these three factors work across the corpora and attempt to interpret them in relation to the properties of individual languages, language change and the role of ELF.

1. Introduction

It is probably fair to say that empirical descriptions of language in use have been for the most part based on average tendencies, rather than preferences of individual speakers. Individual variation has always been recognised, but has seldom been the main focus, except in the fields of stylistic studies and forensic linguistics. In fact, the study of idiolects has been discouraged at least in the field of language change since the seminal publication by Weinreich, Labov and Herzog (1968) who argue that an individual language is not the right place to look for linguistic regularities or change. Labov also clearly stated later that "language […] is an abstract pattern, exterior to the individual" and that "the individual does not exist as a linguistic entity" (Labov 2006: 5).

However, many recent studies indicate a growing interest in the study of variation between individual language users, owing both to the increasing availability of large, diverse and richly annotated datasets and to the finding that these differences do matter in the description of language use. Accordingly, studies have indicated substantial individual differences in grammar (Dąbrowska 2012), collocational preferences (Mollin 2009), n-gram

profiles (Barlow 2013; Wright 2017) and lexico-grammatical patterns (Hall et al. 2017). Vetchinnikova (2017) explicitly sets language representation at the individual level against the communal level, at which language is normally described using data aggregated from a population of individuals, and argues that they can be qualitatively different from each other, in the same way as different dialects of a language are both different from each other and from the 'standard'.

Indeed, if, following a recurring line of thinking in this volume, language (or to be more precise, its communal level) is conceptualised as a complex adaptive system (Ellis and Larsen-Freeman 2006; Beckner et al. 2009; see also Schneider and Mauranen this volume), it can be defined as "a set of variables that interact over time" (de Bot et al. 2007: 7). There is no unanimous agreement on what these interacting variables or elements are. Some researchers like Schneider (this volume) see language as a system of linguistic units at various levels of abstraction, such as phonemes, morphemes or lexemes, entering into syntagmatic and paradigmatic relations with each other. However, it is also possible to see language as a system of interacting idiolects. We are not the first to suggest this view. For example in their position paper, Beckner et al. postulate that language seen as a complex adaptive system "consists of multiple agents (the speakers in the speech community) interacting with one another" (2009: 2). We think these views are not mutually exclusive and depend on the specific time-scale and level of representation being modeled. If the latter approach is adopted, the relationship between the communal average and the individual languages must be characterised by the property of emergence characteristic of complex systems, suggesting that these two levels are qualitatively different. If this holds, substantial changes at the individual representation must bring about change at the communal level.

The issue of individual variation is of particular interest in situations where English is used as a lingua franca (ELF). Typically, in such situations most individual languages are processed as a second language as well as experience other effects of multilingual, rapidly changing environments. As a result, they are likely to become more different from each other, and the description of the average might become less informative than ever before. In fact, variability has been recognised as one of the key features of ELF use (Mauranen 2012, 2017; Jenkins 2015, papers in Jenkins et al. 2017). If such variability, at least in part, is brought about by larger differences between individual languages, the description of their properties might bring in new insights into the study of ELF, changing English and language change more generally.

This paper contributes to the description of the properties of individual languages and their relationship to the communal average. More specifically, we take a linguistic variable, in this case the use of a contracted vs uncontracted form *it's* vs *it is*, and look at the distribution of its variants and factors influencing the choice in individual and communal corpora collected from an online ELF environment. The factors we examine are: SYNTACTIC STRUCTURE, PRIMING and CHUNKING. Our aim is to investigate whether and how these factors work at the individual and communal levels, and in which way they can have an effect on language change.

2. English contraction: Morphosyntactic variation, reduction and chunking

Why do we sometimes use contracted forms, and sometimes uncontracted? There is a number of social, cognitive and linguistic factors which can influence the choice as well as a variety of frameworks within which they are explored. In this chapter, we use the term *contraction* to refer to the process of reducing the expression *it is* to *it's* as well as to other similar predictable reductions.[1] In our written data, the reduction is orthographically marked with an apostrophe, and the form *it's* corresponds to a phonologically reduced form in spoken language.

The choice between full and contracted forms can be looked at as a case of morphosyntactic variation, together with other linguistic variables such as alternation in the dative (*gave a book to him/ gave him a book),* the genitive (*the girl's eyes/the eyes of the girl*), the comparative (*happier/more happy*), relative pronouns (*that/who*) and particle placement (*switch off the light / switch the light off* (see e.g. D'Arcy and Tagliamonte 2015 and Gries 2017 for an overview). Since Labov's seminal work (e.g. Labov 1969), morphosyntactic variables like these have been extremely popular in sociolinguistic variationist analyses, given that it is relatively straightforward to determine variable context, or "different ways of saying the same thing" (e.g. Tagliamonte 2006: 71). The aim of this line of research is "to understand the mechanisms which link extralinguistic phenomena (the social and cultural) with patterned linguistic heterogeneity (the internal, variable, system of language)" (Tagliamonte 2011: xiv; Sankoff 1988: 157). The main assumption here is that linguistic variation reflects social organisation and by correlating linguistic variables as

[1] An example of a non-predictable reduction is *won't*, which Huddleston and Pullum (2002: 91) analyse as an inflectional form of *will.*

dependent variables with social factors as independent variables, one can uncover properties of social structure.

Contraction can also be viewed as a case of morphosyntactic or phonological reduction along with, for example, zero complementation in *that*-clauses (e.g. Jaeger 2006) or word-final /t/ and /d/ deletion (Labov 1972; Bybee 2002). In previous studies, reduction has been explained by three interpretations related to frequency effects, which are cognitive in nature, including 1) chunking and ensuing language change; 2) other frequency effects and 3) rational striving for uniform information density (UID). These explanations are partly overlapping but also partly distinct. In what follows, we explore previous accounts on the relationship of these factors and contraction.

Phonological reduction as a result of chunking has been extensively discussed by Bybee (e.g. 2002, 2010). For example, Bybee and Scheibman (1999) show that the degree of phonological reduction in *don't* is associated with frequent contexts of use, more specifically when *don't* occurs after the first person singular pronoun *I* and before high-frequency verbs such as *know, think, have, mean,* and *feel.* They argue that through chunking, frequent word combinations become processing units with autonomous storage and undergo changes in their constituent structure. Structural changes also couple with changes in meaning, or more precisely, pragmatic functions, and these processes have also been described in several corpus studies in the neo-Firthian tradition (Sinclair 1991; Hunston 2007; Cheng et al. 2009). In all cases where *don't* is phonologically reduced the most, the 'hosting' unit has a function which is different from the same combination of words where *don't* is not reduced. For example, all occurrences of *I don't know* in their data are associated with the literal meaning of 'not knowing' but only the reduced instances convey an additional pragmatic function of speaker uncertainty and mitigation of polite disagreement (Bybee and Scheibman 1999: 587). For Bybee and Scheibman, it is the same change of constituent structure which occurs in grammaticalization.

While chunking allegedly starts off as a result of frequency, frequency effects on morphosyntactic variation have also been examined separately from the notion of chunking, as they seem to be able to have an online, synchronic influence on variation too. What sets the two accounts apart is the fact that the frequency effect account does not presuppose emergence of a unit, which in turn is indispensable in the chunking account. For example, in their study of contraction, Bresnan and Spencer (n.d.) explicitly make a distinction between frequency effects on morphosyntactic variation in compositional and non-compositional

language processing. They find that the effect of joint probability of a lexical subject (i.e. excluding pronoun subjects) plus *is/'s* is larger than the effect of other predictors found in previous studies.

Previous corpus studies on morphosyntactic alternation have used a number of other frequency-based predictors, including the joint probability of the preceding word plus the target, and the joint probability the target plus the following word (e.g. Barth and Kapatsinski 2017). Another measure is conditional probability, that is, the probability of the target given the previous word(s) or the following word(s). Higher probability of the target given its context is expected to lead to reduction (e.g. Jurafsky et al. 2001). Also surprisal has been used in some studies (Wulff et al. 2018).

Another approach using conditional probabilities is the Uniform Information Density (UID) hypothesis (Jaeger 2006), according to which a speaker produces language rationally maintaining a uniform level of information load or density across constructions s/he employs. The UID hypothesis predicts that "elements with high information are lengthened, and elements with low information are shortened", e.g. contracted such as in *you are* > *you're* (Frank and Jaeger 2008: 939). Information density is defined as the conditional probability of the focus element (contracted/uncontracted forms) given the words surrounding it: the more predictable an element is, the less information it contains.

Another factor which works at the cognitive level and is starting to attract more attention is priming (for an early account, see Poplack 1980). In a recent volume edited by Hundt et al. (2017), Pickering and Garrod (2017) and Mair (2017) engaged in a detailed discussion on how cognitive research on priming can be integrated with linguistic and in particular corpus linguistic research on language change. Pickering and Garrod (2017, see also 2004) define priming as "a largely non-conscious or automatic tendency to repeat what one has comprehended or produced" (2017: 173). In their account, priming works towards alignment of interlocutors at different levels of linguistic representation enhancing their mutual understanding. In addition, it contributes to routinisation, or the development of fixed expressions with specific meanings which can start out as *ad hoc* but become conventional over time. Priming is usually studied at very short time-scales, e.g. within a conversation, but can clearly have longer-term effects (e.g. over a week, Kaschak et al. 2011) and, as Pickering and Garrod (2017) argue, possibly lead to permanent changes. Priming is found in adults, children and non-native speakers, and it works at different levels of linguistic representation as well as across them. Syntactic priming, or the tendency to repeat the structure of the

utterance just comprehended or produced, is very common: for example, passives prime passives (Bernolet et al. 2009), even cross-linguistically (Hartsuiker et al. 2004). Importantly, ungrammatical structures can also prime leading to increased acceptability after exposure (Kaschak and Glenberg 2004; Luka and Barsalou 2005). Syntactic priming can be enhanced by a *lexical boost* when the constructions priming each other share the specific verb (Branigan et al. 2000) and when the interlocutor is an actual addressee rather than just a passive listener (Braningan et al. 2007). It also appears stronger for infrequent constructions, possibly because surprising forms are learned better (Bernolet and Hartsuiker 2010).

Other corpus studies on priming effects include Szmrecsanyi (2006), who used the term *persistence*, Barth and Kapatsinski (2017) and Mair (2017), who found that the occurrence of *wanna* (<*want to*) can be primed by previous occurrences of *wanna* and *gonna* in the spoken part of the Corpus of Contemporary American English. Gries (2017) suggests using different priming related factors as autocorrelation effects in multifactorial designs.

In sum, if we leave aside the specific research questions or hypotheses, previous studies have suggested four types of determinants of variation with respect to English contraction: (1) **phonological**, such as preceding segment phonology (Labov 1969; Barth and Kapatsinski 2017; MacKenzie 2012) or, more rarely, rhythmic and segment alternation (Gries 2017), (2) **syntactic**, including different forms of BE – copula, future, progressive and passive (Barth and Kapatsinski 2017) or auxiliary and copula uses (Bresnan and Spencer n.d.) — and the occurrence of the following constituent, often mixing categories from different levels of language organisation[2] (e.g. MacKenzie 2012); (3) **lexical or distributional**, which can subsume a variety of factors based on frequencies of words and their combinations, including subject type (pronoun vs full NP) and length of subject, and, finally, 4) **priming**. Still, there is no unanimous agreement on the factors which determine the choice between full and contracted forms and especially their strength. For Barth and Kapatsinski (2017), this may be due to language redundancy and collinearity among the potential predictors, a point we will discuss in the next section.

3. Individual variation

Individual variation as one of the possible factors affecting the choice is raised in some papers (Gries 2017; Barth and Kapatsinki 2017; see also a discussion in MacKenzie 2012,

[2] For example, MacKenzie (2012) used the following categories, mostly following Labov (1969): adjective, *going to* or *gonna,* quotative *like,* locative *(*e.g. *at work),* noun phrase or clause*,* progressive verb and not available for coding.

Section 1.2.3). However, usually the aim is to ensure that individual preferences do not skew the interpretation of central tendencies, rather than actually focus on them. The individual is sometimes included in mixed-effects models as a random effect, because including more than one data point from the same individual violates the assumption of independence. This approach is also taken in Barth and Kapatsinski (2017), whose point of departure is redundancy in language. Indeed, redundancy is what makes language robust, as meaning is inferable from multiple structural layers at the same time, and understanding meaning works through prediction and confirmation of the predicted. Thus redundancy also means predictability at different levels of abstraction, or structural layers. As Barth and Kapatsinski (2017: 203) put it, "every feature is predictable from multiple other features", and for this reason they use multimodel inference for inferring grammar from a corpus, instead of the more traditional model selection approach (e.g. Labov 1969). They test a set of models which combine different predictors of contracted/uncontracted BE in different ways and find that there is a number of models which perform almost equally well.

We suggest that part of language redundancy at the communal level and the ensuing difficulty of determining the best predictors for a feature can come from individual variation. Usage-based thinking postulates that language with its characteristic structure arises from the interaction of input and application of domain-general cognitive properties, such as categorization, analogy, chunking and prediction (e.g. Bybee 2010). Both input and cognition are individual, which, in principle, should lead to an individual version of the language. At the same time, there must be enough overlap between such individual versions to enable social interaction. Presumably, the discrepancy does not result in communicative problems partly due to shallow or approximate processing, as the differences between individual versions simply go unnoticed. Redundancy serves as the other part of the 'safety net': an individual can process language based on his/her own version of the grammar and it will still work. Divjak and Arppe (2013), who study how categorization works and how prototypes can be obtained from near-synonymous exemplars, also mention redundancy: it seems individual learners can pick up very different combinations of synonym properties from exposure and as a result end up with very different prototypes. This, as they put it, "would make it irrelevant what learners track, as long as they track something" (2013: 245).

If we build our individual versions of the grammar, such grammars might be more consistent (cf. Nevalainen this volume, especially Sections 3.1.3 and 4.1) than an aggregated one. In other words, another way of trying to disentangle language redundancy at the

communal level is to divide up a corpus into individual corpora and examine various predictors per each individual instead of in the aggregate. So while Barth and Kapatsinski (2017: 204) suggest that "the grammar we induce from a corpus is better thought of as an ensemble of models rather than a single model", we suggest that this might be an ensemble of individual grammars.

4. Data

In contrast to many previous studies studying morphosyntactic variation and reduction in spoken data, our focus is on how this phenomenon is manifested in writing, although in a register that resembles spoken language in many ways: blog comments. We use a corpus which contains ca. 7 million words of comments posted on one blog by over 4,000 individual native and non-native English speakers over a period of seven years (the actual blog posts are not included in the data). We focus on 10 most prolific native and non-native commenters on the blog and extract their individual outputs ranging between 40,000 and 246,000 words posted, including the author of the blog (Josef) who contributed ca. 2m words in comments. The structure of our corpus is summarised in Table 8.1. As can be seen, the data includes four American commenters, one Canadian, two Czech, one Greek, one Swiss and one French, thus forming a typical ELF environment. In addition, we treat comments by occasional contributors to the blog as representing the communal average, or the communal level. We call this subcorpus *Non24*, as it is collected from commenters outside the top 24 contributors in terms of volume of output (< 400 comments per commenter).

Table 8.1. Individual subcorpora and Non24.

Commenter	N of tokens	N of years active	NS/NNS	Nationality
Josef	1,752,331	8	NNS	Czech
Gary	246,468	8	NS	US
Carol	231,316	8	NS	US
Louis	183,629	3	NS	US
Graham	174,903	8	NS	Canadian
Ruth	160,161	7	NS	US
Agnes	92,861	3	NNS	Greek
Marek	66,950	6	NNS	Czech, lives in Austria
David	41,111	4	NNS	Swiss, lives in Germany
Sabine	39,963	3	NNS	French, lives in Ireland
Non24	3,549,185	8	Both	ca. 4000 commenters

It is important to point out that the Non24 corpus serves as a reference corpus at the communal level. In other words, it is not conceptualized as representative of the whole corpus, or of the genre of research blogs in general; instead, it is simply a sample of language matched to the individual corpora we focus on in terms of time, genre, sociolinguistic context and discursive situation. The only difference from individual corpora is the fact that it is a sample of language of a large number of individuals where no single individual is overrepresented. Most corpus studies rely on such corpora, which are balanced in terms of individual speakers. Thus, by comparing individual corpora to Non24, we aim to show whether individual languages differ from what we generally know about language based on corpus studies. One of the goals of this research then is to examine whether the exploration of individual corpora is worth pursuing in the future and which research questions such exploration seems to suggest.

Using comments written on one blog as data has a number of advantages. In particular, the genre and social context are constant across individuals, so if contacted/uncontracted form is a purely stylistic choice, it should be categorical within one commenter. Also, the genre of blog writing has been described as situated somewhere between written and spoken modes (e.g. Mauranen 2013; Myers 2009), which makes it a convenient proxy for spontaneous language use similar to much of spoken language, even if it technically represents written language. This is even more true for blog comment threads, which are interactive and less susceptible to norms from above of more formal writing. Such data can even be relevant for reduction accounts, as phonological processing is shown to be at least in some way present in writing and reading (e.g. implicit prosody, Fodor 2002).

5. Methods

Our focus in this study is the use of contracted vs. uncontracted forms in one specific construction: *it is* vs. *it's*. The decision to concentrate on this form, rather than include all forms which allow contraction, was motivated by our specific aims of focusing on individual-based variation, as well as assessing the effect of chunking and priming. This aim is easier to achieve by focusing on a single construction, as previous studies have indeed established that the contracted/uncontracted choice depends on the host of the verb (e.g. Labov 1969; MacKenzie 2012). We chose to focus on *it is*/*it's*, because it is both frequent, providing enough data to allow the analysis at the level of individual speakers, and grammatically versatile, occurring in almost all types of information packaging constructions as categorized

in Huddleston and Pullum (2002). This exceptional property enables us to undertake a more delicate syntactic analysis and extend the earlier investigations of the syntactic factor in morphosyntactic variation.

Accordingly, our analysis first compares the general tendencies in the use of the variants at both the communal and the individual level, and then moves on to investigating three factors which potentially have an effect on the choice between the variants: SYNTACTIC STRUCTURE, PRIMING and CHUNKING, both at the communal level and within each individual. In what follows, we will describe the procedures undertaken to examine each of the factors one by one.

First, we extracted all instances of *it is* and *it's* from 10 individual subcorpora. In addition, we included 2,000 randomly selected instances from *Non24*. Then we carefully checked this initial data set and removed all false positives, including cases where the writer had intended to use a genitive form (e.g ***it's*** [sic] *major economic partner is China*) or where the contracted form *it's* stood for *it has.* We also removed all instances where the word *it* is a prepositional complement and not the head of subject NP (e.g. *the chaotic nature of* ***it is*** *inseparable*), and cases where contraction would not be possible due to the primary verb *is* occurring at the clause-final position (e.g. *I don't know what it is*) (Biber et al. 1999: 1028). We excluded negative forms *it isn't / it's not / it is not*, because they include a choice between three rather than two alternative forms. Our final data set contained 17,994 instances, which we then classified with respect to a number of independent variables:

- PERSON: Possible values include 10 individual commenters and *Non24*.
- PRIMING: The variable is operationalised here as the occurrence of another contracted form in the previous context of 10 words. Possible values are *yes* and *no*.
- CHUNKING: The variable indicates whether *it is*/*it's* is part of chunk, which is operationalised here as a semi-fixed 5-word[3] n-gram (e.g. ***it is*** *a matter of*, ***it's*** *fair to say*). Importantly, we did not generate a general reference list of n-grams to which individual languages would be compared, but created a list of n-grams individually for each speaker. Possible values are *yes* and *no*.
- SYNTACTIC STRUCTURE: Possible values are *cleft*, *progressive, passive*, *extraposition* and *copular.*[4]

[3] The string *it's* is treated here as consisting of two words.
[4] We have excluded from the plots instances of the construction *going to V* due to small number of instances (n=57).

While previous studies have categorised uses of BE into *copular* and *auxiliary*, sometimes using the sub-categories *passive*, *progressive* and *future*, we adopted a more fine-grained categorisation to better reflect the syntactic differences of the constructions in which *it is/it's* is used. One reason for this is that we wanted to avoid overlap with the variable CHUNKING, which is lexical at its core. Thus, we operationalised syntactic structure in an iterative fashion, using Huddleston and Pullum (2002) as the basis and making distinctions between different information-packaging constructions. While clefts and extraposition constructions also represent copular uses of BE, we categorised them separately, especially because copular uses form the majority of our data. Thus, the category *copular* only contains "canonical information-packaging constructions", using Huddleston and Pullum's (2002: Chapter 16) terminology. These copular uses were further categorised based on the type of the predicative complement into *adjective phrases, noun phrases*, *prepositional phrases*, *adverbial phrases* and *clauses.*[5] Extraposition constructions were further divided into *infinitival clauses*, *declarative content clauses*, *interrogative clauses*, *noun phrases* and *gerunds*, based on the type of the extraposed subject. See Table 8.5 in Appendix 8.1 for more details and examples.

In addition, we identified two more categories, which create a syntactic context where the choice between the full and the contracted form is skewed towards the full form. In the first category, a predicative complement precedes *it is/it's* i.e. takes a prenuclear position. This often happens in fused relatives, open interrogatives and relativisation of predicative complement (see Table 8.4 in Appendix 8.1 for examples). In the second, *it is/it's* participates in a comparative construction, i.e. a complement expressing the second term in the comparison which is often elliptical and omits the predicative complement.

Due to scarcity of data in some of the categories, some questions examined are exploratory in nature.

6. Results

6.1. Variation between individuals

The distribution of *it is* and *it's* in our data underlines the importance of looking at the data not only at the communal level, but also at the level of individuals, as these two perspectives exhibit considerable differences. At the aggregate level (Non24), the data shows a moderate preference for the full form, with with 39% (648) of the total instances having the form *it's*

[5] Clauses include *fused relatives, content clauses* and *infinitival clauses.*

and 61% (1016) the form *it is*. However, individual commenters vary greatly between their preferences, as can be seen in Figure 8.1, where the speakers are ordered by the percentage of the full form (shown on the y-axis) from left to right.

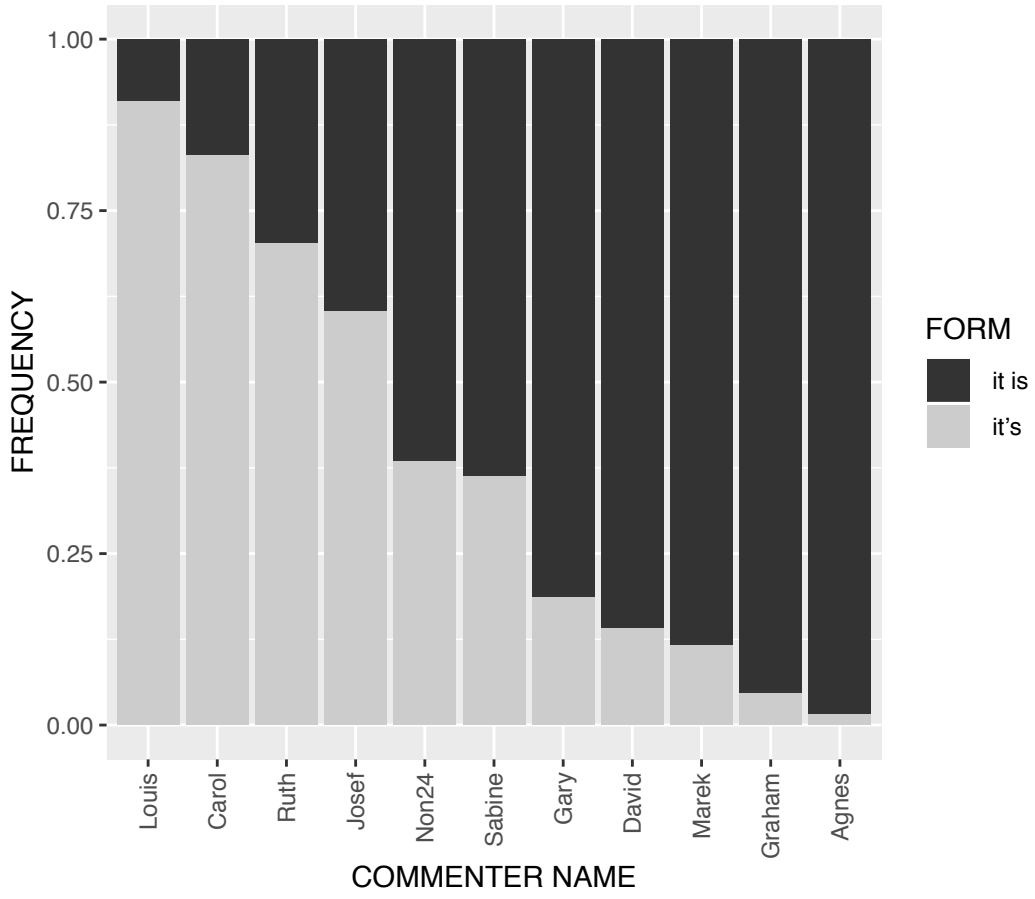

Figure 8.1. Proportions of *it's* and *it is* per PERSON.

Most individual commenters have an overwhelming preference for either the full form (Agnes, Graham, Marek, David and Gary) or the contracted form (Ruth, Carol and Louis). Only Sabine and Josef, the two remaining commenters in the middle, do not show a clear preference towards either variant. It is thus clear that what looks like as variation in the aggregate can be heavily partitioned at the individual level. From this follows that for some commenters, there is no possibility for any other factors to further influence the choice: for example Agnes and Graham are categorical in their preference for the full form, which can be a stylistic choice in that it is unaffected by any changes in the conditions. Next, we shall look more closely at the remaining data sets where the influence of other factors can still be considered.

6.2. Syntactic structure: the main categories

Contrary to what could perhaps be expected, SYNTACTIC STRUCTURE does not appear to have a systematic influence on the choice between the two forms. As can be seen in Figure 8.2, there are only very small variations in the proportions across all syntactic structures in all data sets. When subcategories of the largest categories, *copular* and *extraposition,* are considered, the overall picture stays the same. There are some effects visible, but they vary both in direction of the influence and the syntactic category for different individuals.[6]

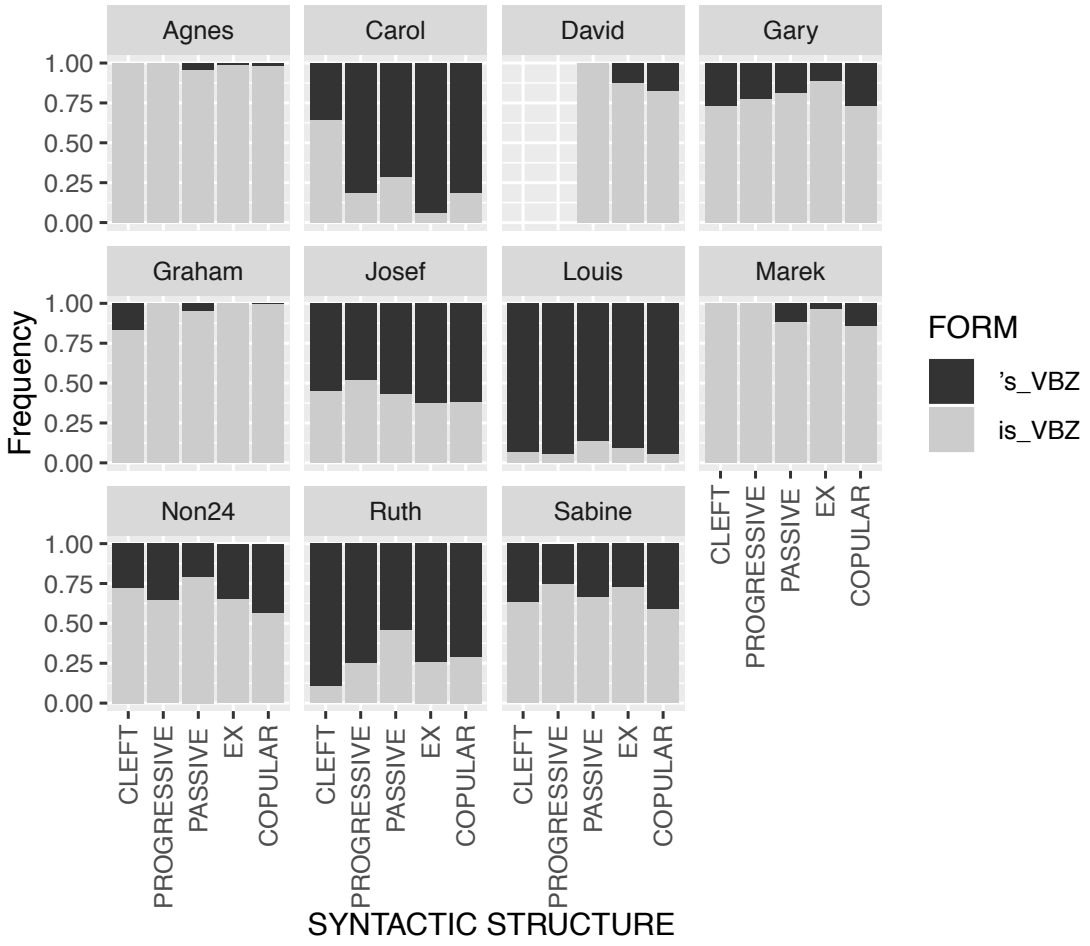

Figure 8.2. Proportions of *it's* and *it is* per SYNTACTIC STRUCTURE, PERSON.

[6] As mentioned in Section 5, we did not include *Going to V* category in the plots due to low frequency. There are only 57 occurrences of the construction in the whole data set, out of which 23 belong to Louis's subcorpus (1,3 instances per 10k words). Agnes, David, Graham and Sabine do not use the construction at all and the rest use it only very rarely (less than 0,3 instances per 10k words). In Non24, the construction occurs seven times, five of which are contracted and two uncontracted. In contrast, all instances in Louis's corpus are contracted. This distribution also suggests individual preferences.

At the communal level (Non24), the contracted form seems to be positively associated with the *copular*[7] category and negatively associated with the *passive* (χ^2=26.7, df=4, p<0.05). However, these associations are not observed in individual corpora. Yet if we look at the subcategories of *copular* constructions, we see that when BE is followed by a *clause*, all commenters who prefer the contracted form overall show an even stronger preference for contraction, whereas those preferring full forms are unaffected (Agnes, Graham and Marek, but with the exception of Gary).[8] It is possible that we see an effect of chunking here, as structures where *it* functions as subject, BE as copula and clause as complement tend to be idiomatic (Huddleston and Pullum 2002: 962). This is especially true for content clauses, such as (1) – (3):

(1) ***It's*** *just that you haven't really read the paper…* (Josef)
(2) ***It's*** *just that the juxtaposition doesn't work…* (Louis)
(3) *I think* ***it's*** *because they refuse to believe that…* (Carol)

Here, *it's just that* and *it's because* seem to work as semi-fixed chunks, sometimes allowing a modifier (e.g. *it's always because*), which serve as convenient sentence openers.

Two more trends, this time completely idiosyncratic, can be seen in Carol's usage. First, her overall preference for the contracted form is overruled in clefts where she uses the full form 18 times and the contracted one 10 times. No other commenter with a preference for contracted forms shows any tendency in this direction: Louis and Ruth are almost categorical in using the contracted form in this category too. In contrast, Carol's overall preference for contraction is even more pronounced in the *extraposition* category (including all subcategories). Given that clefts and extraposition constructions belong to the same higher order category of BE used as copula, the explanation for the variation is unlikely to be purely grammatical. Another idiosyncratic trend is observable in Josef' data: he prefers to use the full form in progressives despite his overall preference for contraction.

The reasons for a range of trends reported here are not immediately clear. We have mentioned chunking as one of the possible explanations and will return to it in more detail in Section 6.4. What is clear though is that syntax does not have a systematic influence on the

[7] Note that since we distinguished between different information-packaging constructions, the category *copular* only contains uses of BE as copula in canonical syntactic structures. Clefts and extraposition constructions where BE also serves as copula are analysed as separate categories.
[8] The choice of the *it's* in this context is nearly categorical for Carol (10 out of 11 instances), Louis (33/34) and Ruth (12/14), and the probability of the contracted form is considerably higher for Gary (10/19) and Josef (214/252) compared to their general tendencies.

variation between the forms in our data; instead, what we see is mostly a collection of individual preferences. This finding supports the hypothesis that what we see at the communal level can be quite different from what happens at the individual.

6.3. Special syntactic context

We will now briefly comment on two special categories mentioned in Section 5 where syntactic structure influences the choice quite clearly in favour of the full form: cases where a predicative complement takes a prenuclear position (e.g. *to know* ***what*** *it is really about*) and comparative constructions (e.g. *worse than it is now*). While the preference for the full form is almost categorical — and for this reason we were originally going to exclude these instances altogether — cases of contraction also occur, suggesting they should be included in the analysis.

Out of 157 prenuclear cases, in only 22 the verb is contracted. These include 20 instances where *it's* is part of some conventional phrase: *for what it's worth* (12 instances), *what it's like* (5), *where it's at* used in an idiomatic sense (2) and *what it's all about* (1). It therefore seems that the dispreferred contraction here can be partly explained by chunking, which increases the likelihood of reduction. At the same time, all but one occurrence belong to commenters who have an overall preference for contraction: Louis, Carol, Josef and Ruth, indicating a tentative possibility of some interaction between the preference for contraction and use of chunks.

The other category is formed of comparative constructions. In this category, *it's/it is* is part of a structurally reduced subordinate clause expressing the secondary term in the comparison, highlighted in bold in examples (4)–(6) (see Huddleston and Pullum 2002: 1106).[9] As a result of such structural reductions, the predicative complement often does not follow *it's/it is* just like in prenuclear cases. For example:

(4) *This is as true for the Greeks* ***as it is for us****.* (Carol)

(5) *And as painful* ***as it is*** *to say this…* (Louis)

(6) *The higher CO2 is,* ***the harder it is to increase it****…* (Josef)

However not all comparative clauses are elliptical or structurally reduced, and we were interested to see whether the preference for the full form due to the absence of the complement in its habitual position spreads or generalises to cases where the complement

[9] For example, the sentence **she is older than I am old* is ungrammatical (Huddleston and Pullum 2002: 1108).

follows *it's/it is* as normal and in principle nothing precludes the use of contraction (e.g. *My explanation is as elegant as it is simple…(*Non24)). [10] Out of 180 cases of the comparative construction, 108 are prenuclear, and the full form was used in all of them. This finding supports the effect of the prenuclear position of the predicative complement on the variation between the contracted and the full form. Out of the remaining 72 cases where the predicative complement was present and positioned after *it's/it is,* 45 belong to Josef, which only allows us to look at the variation in his data set. In this syntactic context, Josef used the contracted form 6 times, and the full form 39 times including 24 cases involving a reanalysed construction *as long/far as* functioning as a compound preposition (see Huddleston and Pullum 2002: 1134). What is interesting is that this uncharacteristic preference for the full form is also attested here. Whether Josef's preference for the full form in all comparative clauses is another case of an individual preference, or whether it is motivated by generalisation from prenuclear cases, remains an open question. If the latter is true, the process of generalisation can also be studied at the individual level.

In general, it seems while overall syntax does not have a systematic influence on the variation, special syntactic conditions can set constraints on the use of the contracted and uncontracted forms.

6.4. Priming

Next, we investigated the effect of PRIMING, more precisely the possibility that the choice of the contracted form would be primed by a previous use of contraction within the same stretch of discourse (Szmrecsanyi 2006; Barth and Kapatsinski 2017; Mair 2017). As previously described, priming was operationalised as the preceding occurrence of another contracted form within a window of ten words, similar to Barth and Kapatsinski (2017) and Mair (2017). Figure 8.3 shows the proportions of contracted and uncontracted forms separately for cases where there is not a previous occurrence of a contraction ("NO", left bar) and where there is one ("YES", right bar). The left bar only includes instances where there would have been an opportunity for using a contracted form; other instances have been excluded.

[10] See Barth and Kapatsinski (2017: 249) for a discussion of a similar issue of the direction of fusion with the preceding word generalising beyond contexts motivated by co-occurrence: *it's* vs *cat's*.

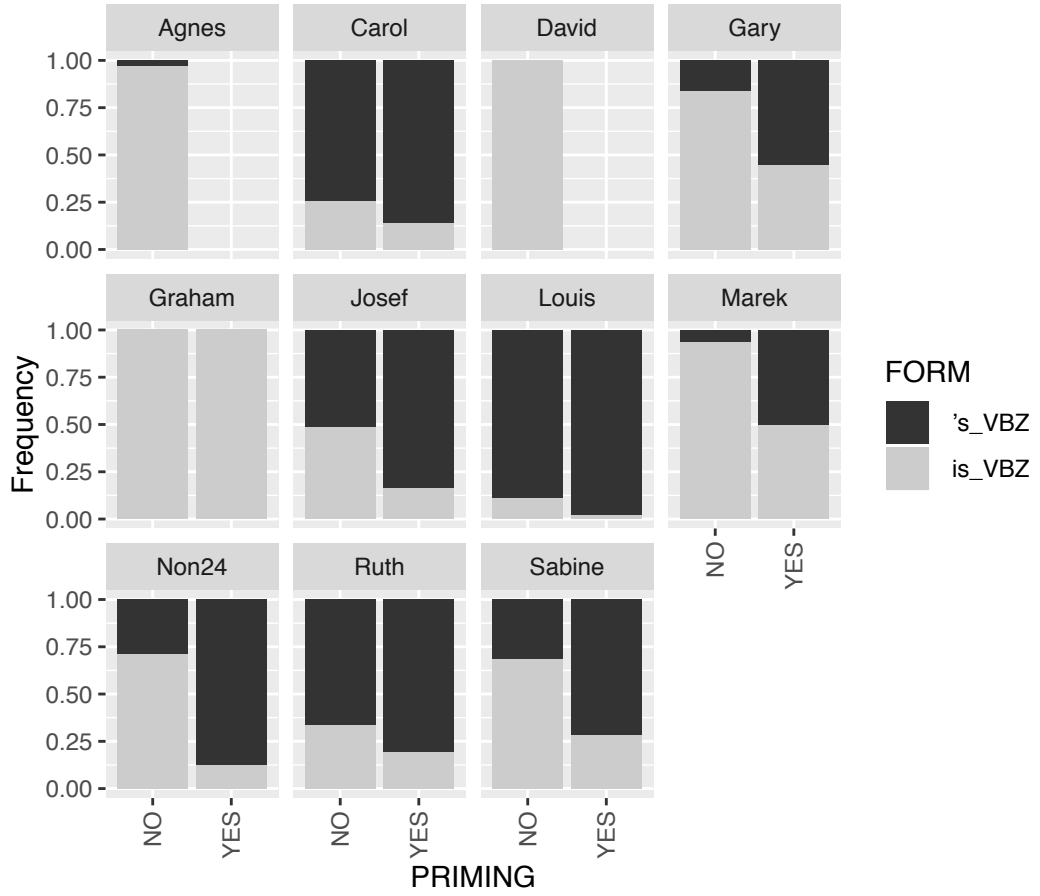

Figure 8.3. Proportions of *it's* and *it is* per PRIMING, grouped by PERSON.

As can be seen, the presence of a previous instance of a contracted form clearly increases the likelihood of using a contracted form. The magnitude of this effect varies between speakers: it is smaller among individuals who prefer to contract and larger among those who prefer to use the full form. This is probably because those who contract overall contract very often even when there is no priming context. Still, even for these individuals the tendency to contract is boosted in the presence of the priming context in comparison to their overall proportion of the contracted forms.

Interestingly, the greatest difference between priming and no priming contexts is found at the communal level, i.e. Non24 data (nearly 60 percentage point). It is difficult to say why this happens, but one possible explanation is that our operationalisation of the priming effect (i.e. previous occurrence of a contracted form or a full form within ten words to the left) not only identifies the priming contexts but also splits the data into instances produced by those who prefer to contract and those who prefer to use the full form, as Non24 data is very likely to contain both types. In other words, if we find two instances of the contracted form within the span of 10 words, this can happen 1) due to priming or 2) because both instances are

produced by the same individual who prefers to contract irrespective of priming. If this is indeed the case, studies which use communal data might be exaggerating the effect of priming (in the case of self-priming): one speaker can be using the same form because s/he has a preference for this form overall, rather than due to online priming.

More generally though, the direction of the effect we observe is the same for all speakers, which strongly suggests that priming has a systematic effect on the variable in focus.

6.5. Chunking

The final variable we look at is whether *it is/it's* is part of a chunk in the use of a specific individual, referred to here as CHUNKING. With more data it should be possible to look at degrees of constituency, but here we operationalise the variable only in a dichotomous way as compositional vs non-compositional processing. We also take a very conservative measure of non-compositionality, counting a sequence of words as a chunk if it is a fixed 5-gram and occurs in an individual corpus at least three times. The only variation we allow is the variation between *it is/it's.* 5-grams are comparatively rare even in large corpora and if they recur in a relatively small individual corpus, compositional processing is highly unlikely. However, the conservativeness of the measure needs to be taken into account in interpreting the results.

The fact that we extract 5-grams from each corpus individually means that in our view chunks can be personal. In other words, we assume that a sequence of words which is processed compositionally by the majority of the population (the communal level), can be processed non-compositionally by a specific individual and vice versa. Thus, we do not compare individual production against a list of multi-word units established at the communal level and instead create an individual list of chunks using independent criteria, that is, length, fixedness and frequency of repetition. We apply the same criteria of non-compositionality to Non24 data for the purposes of comparison as chunks are normally identified in corpus linguistics using aggregate data.

Figure 8.4 shows the proportions of the variants separately for instances which are not part of a chunk ("NO") and which are part of a chunk ("YES"). Excluded from the figure are commenters whose data does not contain 5-grams that meet the frequency threshold. As can be seen, similar to PRIMING, CHUNKING seems to increase the likelihood of the contracted form across individual speakers. This clearly holds at least for those who prefer to contract in

general.[11] Yet, the effect reaches statistical significance only for Josef and Carol. This can be due to the conservativeness of our operationalisation: since our criteria for identification of chunks targeted precision over recall, while we can be sure that the 'YES' bar contains chunks only, the 'NO' bar is likely to contain some proportion of non-compositional instances too.

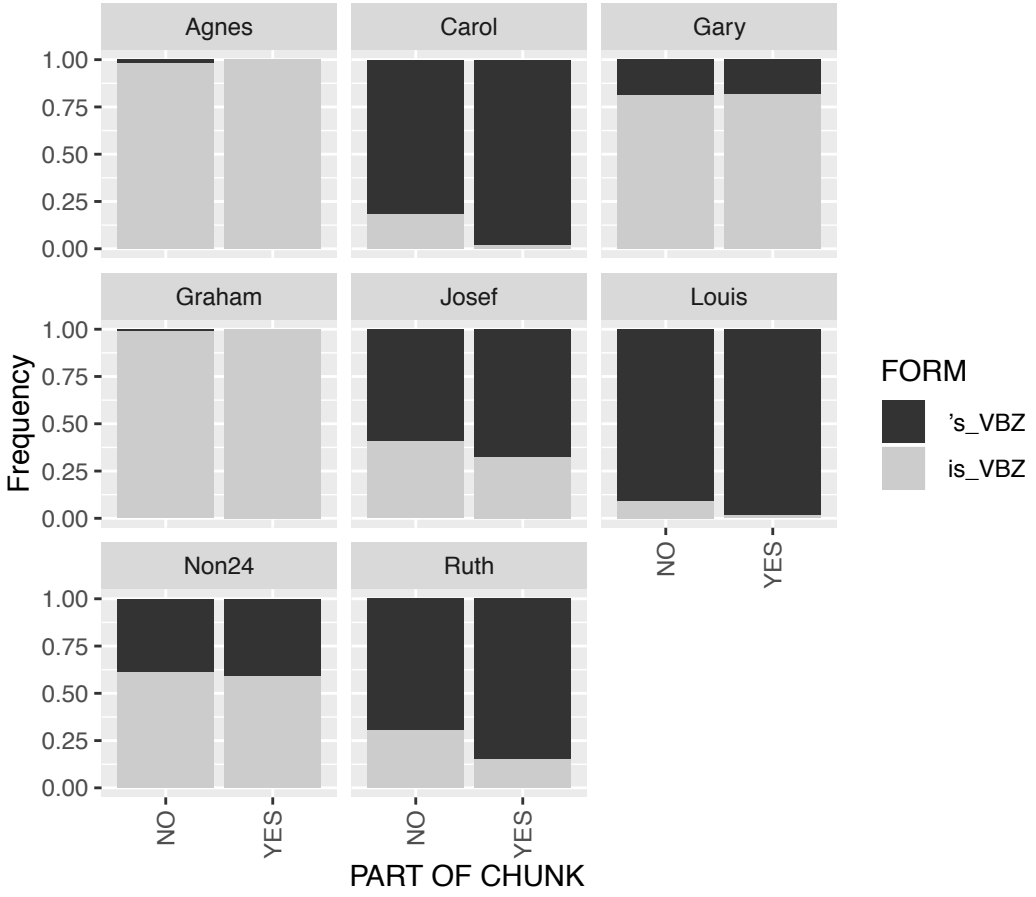

Figure 8.4. Proportions of *it is* and *it's* per CHUNKING, grouped by PERSON.

In principle, the results, at least for Carol and Josef, support the hypothesis that chunking can lead to reduction within speaker-specific chunks. According to a usage-based model of grammar, such reduction presumably occurs due to reanalysis of constituent structure and is therefore a sign of grammatical change (Bybee and Scheibman 1999; Bybee 2010). The individuality of personal chunks can be seen in Table 8.2, which lists four most frequent 5-grams starting with *it's/it is* for each of the commenters.

[11] Gary's data does not support the tendency, Agnes and Graham use full forms categorically irrespective of any possible additional factors as was mentioned in Section 6.1 and David's, Sabine's and Marek's data does not contain chunks according to our criteria, most likely due the size of their corpora (they have the smallest) and conservativeness of our operationalisation of chunks.

Table 8.2. Most frequent 5-grams for each of the individual commenters

Carol	Gary	Louis	Josef	Ruth
quite conceivable that, 13	really hard to, 5	going to be, 8	still true that, 104	kind of like, 6
bad enough that, 8	up to the, 5	quite possible that, 5	very clear that, 64	just me, but, 4
simply the case, 6	true that the, 5	a lot easier, 5	clear that the, 29	a matter of, 4
a good thing, 6	also possible that, 5	a lot more, 5	a part of, 29	too bad that, 3

Thus, chunking, just like priming, seems to have a systematic effect on the variation between the two forms across individuals. This is not very surprising since both chunking and priming have always been proposed as cognitive factors. In this sense, the present study simply confirms that indeed we can see the effect at the individual level. However, as Table 8.2 shows, individuals develop very different repertoires of personal chunks. Thus, the effect of chunking results in individuals contracting in different rather than similar places, irrespective of e.g. syntactic structure.

In fact, it can be tested whether the fact that we did not see a systematic effect of syntactic structure is actually due to the differences between individual chunk repertoires. Unfortunately our data is too scarce to be tested for statistical significance, but a few tentative observations can be made. In Section 6.2 we reported a few miscellaneous effects of syntactic structure on the variation in individual data sets: Carol was shown to prefer contracted forms overall, yet in her data clefts were associated with a lower than expected frequency of contracted forms. Based on Table 8.3, showing the distribution of their personal chunks across syntactic categories, we could hypothesise that Carol's low frequency of contracted form in *clefts* could be due to the total absence of chunks in this category. Most of her chunks are found in *extrapositions*, which is the category most strongly associated with *it's* in her data. Following a similar logic, Josef's lower than expected rate of contraction in the progressive could tentatively be attributed to the near-absence of chunks.

Table 8.3. Distribution of specific chunks for Carol and Josef.

	Cleft	Copular	Extraposition	Passive	Progressive
Carol	0	9	71	0	7
Josef	30	721	1151	21	2

For example:

(7) *it is always the majority who must be right…* (Carol, cleft)

(8) *it is the neoconservatives, not the paleoconservatives, who are to blame…* (Carol, cleft)

(9) *it is the veterinary researcher who first developed a successful IVF technique…* (Carol, cleft)
(10) *it's quite conceivable that the TV producers fabricated this call-girl tale…* (Carol, Ex)
(11) *it's quite refreshing to see a publication such as…* (Carol, Ex)
(12) *it's simply the case that news jibes well with strings…* (Carol, Ex)
(13) *it's promising the people a fixed amount of products…* (Josef, progressive)
(14) *it is rotating and orbiting and moving in many ways…* (Josef, progressive)

Based on Table 8.3 and examples (7)–(14), it is possible to tentatively hypothesise that a heightened preference for contracted forms is associated with categories which are more conducive to chunk formation (copular and extraposition), whereas the categories less conducive to this (such as clefts, passives and progressives) would exhibit lower rates of contraction. As the examples illustrate, the focus of a cleft as well as the main verb in a progressive structure are always changing, while the opening part in extraposition can be very repetitive. In our data, especially extraposition as a syntactic structure seems to facilitate chunk formation across all speakers. This agrees well with the fact that some of the grammar patterns which belong to the category, such as *it is* ADJ *that* (Francis et al. 1998: 480) are described as useful and frequently used for example in academic writing (Charles 2004; Groom 2005; Hunston 2010; see also Biber et al.1999: 1020 on lexical bundles initiating extraposed structures) and are therefore likely to be common in research blogging and blog comments, which is the register our data represents.

What is slightly puzzling in this light is how Carol manages to have chunks in the progressive category, especially as nobody else has any at all. At closer examination, it turns out that many of the instances in her progressive category are actually extraposed structures with a verb phrase instead of a more usual adjective phrase. Most of the instances are variants of the same semi-fixed chunk: *it's /is becoming increasingly clear/apparent/obvious/more difficult to/that* (n=10), such as:

(15) It's becoming increasingly more difficult to see all those distant points…
(16) It's becoming increasingly apparent that the cosmological constant is finely-tuned.
(17) It's becoming increasingly clear that classical algorithms are maxing out.

Other verb phrases initiating an extraposed structure are also possible, e.g.: *it's looking more and more like this isn't the case*.

All cases considered, chunking seems to play an important role in explaining the variation.

7. Conclusions

Normally when we want to describe language regularities, we collect a balanced sample of language from a large number of speakers making sure that no one is overrepresented. The underlying assumption is that language is homogeneous enough across speakers and we are interested in what they share instead of their idiosyncrasies, as the task of describing idiosyncratic features of all individuals is obviously futile. However, it is also generally agreed that linguistic regularities emerge as a result of the interaction between linguistic, social and cognitive factors. In principle, all these factors should then be observable within each individual too. At the same time, the ubiquity of the cognitive factor does not necessarily presuppose uniformity of linguistic regularities. All individuals have the same cognitive properties: they chunk, use analogy, make categories and generalise to new instances. However, linguistic regularities which arise from the application of these properties crucially depend on the input to which they are applied. To what extent different individuals infer different regularities from the inputs they receive is a question which has not received much attention. It also becomes especially interesting in ELF settings, where variation in individual language exposure is much higher than in monolingual settings.

In this study, we selected a linguistic variable – *it is* vs *it's* – and examined the effects of the linguistic (syntactic structure) and cognitive factors (priming and chunking) on the variation in one communal and ten individual native and non-native corpora extracted from the same blog, that is, keeping the social context constant. Importantly, we thus compared individual data sets to a separate corpus with balanced sampling representing the communal level (Non24) instead of to their average. We found that most individuals in our data either prefer the full form or the contracted form. Thus, what the communal corpus probably shows is which preference is in the majority at the population level in this context.

SYNTACTIC STRUCTURE did not have a systematic effect on the variation across all speakers. However, specific syntactic structures seemed to have an effect on specific individuals. Given the chunking patterns of these individual speakers, it is possible that what looks like a syntactic effect might in fact be driven by chunking

Despite the lack of evidence with respect to the effect of syntactic structure in general, the linguistic factor should not be dismissed, as it can set constraints on the variation which

hold across all speakers. In our data, all individuals irrespective of their overall preference were influenced by the constraining factor of the predicative in the prenuclear position and selected the full form in this syntactic context. There is some initial evidence of generalisation from contexts favouring full forms to other instances of the same category working at the individual level, but more research is needed. In principle, since generalisation is a cognitive property, it should be possible to trace it at this level. It would be interesting to investigate which categories individuals build and how different such categories can be across individuals.

In contrast, both cognitive factors of PRIMING and CHUNKING had a more systematic effect across individual speakers. Priming is highly sporadic and it is not clear whether its effects can accumulate and lead to change. Yet, self-priming might be one of the reasons behind clear individual preferences we observed, along with the fact that they do not seem to become more like each other as usage-based theory would predict, a phenomenon described by Barlow (2013: 471) as "inbuilt inertia" (see also Szmrecsanyi 2006; Pickering and Garrod 2017). In contrast, chunking facilitates recurrence and can thus easily accumulate. At the same time, chunking also leads to the development of individual chunk repertoires, which can be substantially different across speakers. Thus, while chunking is a common factor influencing variation between the two forms, it may result in different people contracting in different places.

We did not find an effect of chunking on contraction in Non24. One possible reason for this is that Non24 did not contain as many chunks as individual data sets. This is in agreement with an earlier finding that individual languages are more "fixed" than the communal average, that is, they contain more verbatim chunks: at the communal level the chunks are likely to include variable slots which prevent them from being retrieved with n-gram tools (Vetchinnkova 2017). A possible link between the preference for contraction and the proportion of chunks in one's language (i.e. its degree of "fixedness" (Vetchinnikova 2017) or "chunkedness" (Arnon and Christiansen 2017; McCauley and Christiansen 2017)) might be another interesting question for future research.

There has been some discussion in ELF literature on whether non-native ELF speakers process language compositionally or not, since they do not always reproduce standard phraseological units precisely (Seidlhofer 2009; Pitzl 2012; Mauranen 2012; Vetchinnikova 2015). In our data, both native and non-native speakers exhibited personal repertoires of chunks and showed the effect of chunking on contraction, which speaks in

favour of non-compositional processing. If so, imprecision or approximation (Mauranen 2005, 2012) introduced by non-native ELF speakers into standard English phraseology can also have a long-term effect and implications for language change. It is also possible to hypothesise that individual repertoires of chunks are overall more different from each other in ELF environments than in e.g. closer-knit, predominantly monolingual communities (cf. Laitinen and Lundberg this volume on social network theory and ELF). However, to test this prediction one would need to compare individual chunk repertoires from samples of ELF and monolingual interactions. We did not pursue this goal here.

Overall, in our study the communal level seemed to be different from the individual. Complexity theory provides a good framework for explaining how the two are related. The communal level is not simply the average of the individual languages: such explanation would question the validity of counting the means which is not in doubt. Instead, the communal level can be seen as emergent from the individual, that is, as qualitatively different from it. This explanation makes both levels important and worth of investigation in their own right.[12]

References

Arnon, Inbal & Morten H. Christiansen. The role of multiword building blocks in explaining L1–L2 differences. *Topics in Cognitive Science* 9(3). 621–636.

Barlow, Michael. 2013. Individual differences and usage-based grammar. *International Journal of Corpus Linguistics* 18(4). 443–478

Barth, Danielle & Vsevolod Kapatsinski. 2017. A multimodel inference approach to categorical variant choice: construction, priming and frequency effects on the choice between full and contracted forms of am, are and is. *Corpus Linguistics and Linguistic Theory* 13(2). 203–260.

Beckner, Clay, Richard Blythe, Joan Bybee, Morten H. Christiansen, William Croft, Nick C. Ellis, John Holland, Jinyun Ke, Diane Larsen-Freeman & Tom Schoenemann. 2009. Language is a complex adaptive system: Position paper. *Language learning* 59(s1). 1–26.

Bernolet, Sarah & Robert J. Hartsuiker. 2010. Does verb bias modulate syntactic priming? *Cognition* 114(3). 455–461.

[12] We would like to thank the author of the blog for generously providing his blog data for research purposes.

Bernolet, Sarah, Robert J. Hartsuiker & Martin J. Pickering. 2009. Persistence of emphasis in language production: A cross-linguistic approach. *Cognition* 112(2). 300–317.

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad & Edward Finegan. 1999. *The Longman grammar of spoken and written English.* London: Longman.

Branigan, Holly P., Martin J. Pickering & Alexandra A. Cleland. 2000. Syntactic co-ordination in dialogue. *Cognition* 75(2). B13–B25.

Branigan, Holly P., Martin J. Pickering, Janet F. McLean & Alexandra A. Cleland. 2007. Syntactic alignment and participant role in dialogue. *Cognition* 104(2). 163–197.

Bresnan, Joan & Jessica Spencer. (unpublished manuscript). *Frequency and variation in English subject-verb contraction*. Stanford, CA: Stanford University Department of Linguistics and Center for the Study of Language and Information ms.

Bybee, Joan & Joanne Scheibman. 1999. The effect of usage on degrees of constituency: The reduction of *don't* in English. *Linguistics* 37(4). 575–596

Bybee, Joan. 2002. Word frequency and context of use in the lexical diffusion of phonetically conditioned sound change. *Language Variation and Change* 14(3). 261–290.

Bybee, Joan. 2006. From usage to grammar: The mind's response to repetition. *Language* 82(4). 711–733.

Charles, Maggie. 2004. The construction of stance: A corpus-based investigation of two contrasting disciplines. Unpublished PhD thesis, University of Birmingham.

Cheng, Winnie, Chris Greaves, John McH. Sinclair & Martin Warren. 2009. Uncovering the extent of the phraseological tendency: Towards a systematic analysis of concgrams. *Applied Linguistics* 30(2). 236–252.

D'Arcy, Alexandra & Sali A. Tagliamonte. 2015. Not always variable: Probing the vernacular grammar. *Language Variation and Change* 27(3). 255–285.

Dąbrowska, Ewa. 2012. Different speakers, different grammars Individual differences in native language attainment. *Linguistic Approaches to Bilingualism* 2(3). 219–253.

de Bot, Kees & Diane Larsen-Freeman. 2013. Researching second language development from a dynamic systems theory perspective. In Marjolijn Verspoor, Kees de Bot & Wander Lowie (eds.), *A dynamic approach to second language development: Methods and techniques*, 5–23. Amsterdam; Philadelphia: John Benjamins.

Divjak, Dagmar & Antti Arppe. 2013. Extracting prototypes from exemplars What can corpus data tell us about concept representation? *Cognitive Linguistics* 24(2). 221–274.

Ellis, Nick C. & Diane Larsen-Freeman. 2006. Language emergence: Implications for Applied Linguistics–Introduction to the special issue. *Applied Linguistics* 27(4). 558–589.

Fodor, Janet D. 2002. Psycholinguistics cannot escape prosody. *Proceedings of Speech Prosody 2002. 83–90. Aix-en-Provence, France.*

Francis, Gill, Susan Hunston & Elizabeth Manning. 1998. *Collins COBUILD Grammar Patterns: Nouns and adjectives*. London: HarperCollins.

Frank, Austin F. & T. Florian Jaeger. 2008. Speaking rationally: Uniform information density as an optimal strategy for language production. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 30. 939-944.

Gries, Stefan Th. 2017. Syntactic alternation research: Taking stock and some suggestions for the future. *Belgian Journal of Linguistics* 31(1). 8–29.

Groom, Nicholas. 2005. Pattern and meaning across genres and disciplines: An exploratory study. *Journal of English for Academic Purposes* 4(3). 257–277.

Hall, Christopher J., Jack Joyce & Chris Robson. 2017. Investigating the lexico-grammatical resources of a non-native user of English: The case of can and could in email requests. *Applied Linguistics Review* 8(1). 35–59.

Hartsuiker, Robert J., Martin J. Pickering & Eline Veltkamp. 2004. Is syntax separate or shared between languages? Cross-linguistic syntactic priming in Spanish-English bilinguals. *Psychological Science* 15(6). 409–414.

Huddleston, Rodney & Geoffrey K. Pullum (eds.). 2002. *The Cambridge grammar of the EnglishlLanguage*. Cambridge: Cambridge University Press.

Hundt, Marianne, Sandra Mollin & Simone E. Pfenninger. 2017. *The changing English language: Psycholinguistic perspectives*. Cambridge: Cambridge University Press.

Hunston, Susan. 2007. Semantic prosody revisited. *International Journal of Corpus Linguistics* 12(2). 249–268.

Hunston, Susan. 2010. Starting with the small words. In Ute Römer & Rainer Schulze (eds.), *Patterns, meaningful units and specialized discourses*, 7-30. Amsterdam; Philadelphia: John Benjamins.

Jaeger, T. Florian. 2006. *Redundancy and syntactic reduction in spontaneous speech*. Unpublished doctoral dissertation, Stanford University.

Jenkins, Jennifer. 2015. Repositioning English and multilingualism in English as a Lingua Franca. *Englishes in Practice* 2(3). 49–85.

Jenkins, Jennifer, Will Baker & Martin Dewey (eds.). 2017. *The Routledge handbook of English as a Lingua Franca* (Routledge Handbooks in Applied Linguistics). Milton Park, Abingdon, Oxon ; New York, NY: Routledge.

Jurafsky, Daniel, Alan Bell, Michelle Gregory & William D. Raymond. 2001. Probabilistic relations between words: Evidence from reduction in lexical production. *Typological studies in language* 45. 229–254.

Kaschak, Michael P. & Arthur M. Glenberg. 2004. This construction needs learned. *Journal of Experimental Psychology: General* 133(3). 450.

Kaschak, Michael P., Timothy J. Kutta & Christopher Schatschneider. 2011. Long-term cumulative structural priming persists for (at least) one week. *Memory & Cognition* 39(3). 381–388.

Labov, William. 1969. Contraction, Deletion, and Inherent Variability of the English Copula. *Language* 45(4). 715–762.

Labov, William. 1972. *Sociolinguistic patterns*. Philadelphia: University of Pennsylvania Press.

Labov, William. 2006. *The social stratification of English in New York City*, 2nd edn. Cambridge: Cambridge University Press.

Luka, Barbara J. & Lawrence W. Barsalou. 2005. Structural facilitation: Mere exposure effects for grammatical acceptability as evidence for syntactic priming in comprehension. *Journal of Memory and Language* 52(3). 436–459.

MacKenzie, Laurel E. 2012. *Locating variation above the phonology*. University of Pennsylvania PhD Thesis.

Mair, Christian. 2017. From priming and processing to frequency effects and grammaticalization? Contracted semi-modals in present-day English. In Marianne Hundt, Sandra Mollin & Simone E. Pfenninger (eds.), *The changing English language*, 191–212. Cambridge: Cambridge University Press.

Mauranen, Anna. 2005. English as a Lingua Franca—an unknown language? In Giuseppina Cortese & Anna Duszak (eds.)*, Identity, community, discourse: English in intercultural settings*, 269–293. Frankfurt: Peter Lang.

Mauranen, Anna. 2012. *Exploring ELF: Academic English shaped by non-native speakers*. Cambridge: Cambridge University Press.

Mauranen, Anna. 2013. Hybridism, edutainment, and doubt: Science blogging finding its feet. *Nordic Journal of English Studies* 12(1). 7-36.

Mauranen, Anna. 2017. Conceptualising ELF. In Jenkins, Jennifer, Will Baker & Martin Dewey (eds.), *The Routledge handbook of English as a Lingua Franca* (Routledge Handbooks in Applied Linguistics), 7-24. Milton Park, Abingdon, Oxon ; New York, NY: Routledge.

McCauley, Stewart M. & Morten H. Christiansen. 2017. Computational Investigations of Multiword Chunks in Language Learning. *Topics in Cognitive Science* 9(3). 637–652.

Mollin, Sandra. 2009. "I entirely understand" is a Blairism: The methodology of identifying idiolectal collocations. *International Journal of Corpus Linguistics* 14(3). 367–392.

Myers, Greg. 2009. *The discourse of blogs and wikis*. London: Continuum

Pickering, Martin J. & Simon Garrod. 2004. Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences* 27(02). 169–190.

Pickering, Martin J. & Simon Garrod. 2017. Priming and language change. In Marianne Hundt, Sandra Mollin & Simone E. Pfenninger (eds.), *The changing English language*, 173–190. Cambridge: Cambridge University Press.

Pitzl, Marie-Luise. 2012. Creativity meets convention: idiom variation and re-metaphorization in ELF. *Journal of English as a Lingua Franca* 1(1). 27-55.

Poplack, Shana. 1980. The notion of the plural in Puerto Rican English: Competing constraints on (s) deletion. In William Labov (ed.), *Locating language in time and space*, 55–67. New York: Academic Press.

Sankoff, David. 1988. Sociolinguistics and syntactic variation. In Frederick J. Newmeyer (ed.), *Linguistics: The Cambridge survey,* 140–161. Cambridge: Cambridge University Press.

Seidlhofer, Barbara. 2009. Accommodation and the idiom principle in English as a Lingua Franca. *Intercultural Pragmatics* 6(2). 195–215.

Sinclair, John McH. 1991. *Corpus, concordance, collocation*. Oxford: Oxford University Press.

Szmrecsanyi, Benedikt. 2006. *Morphosyntactic persistence in spoken English: A corpus study at the intersection of variationist sociolinguistics, psycholinguistics, and discourse analysis.* Berlin: Mouton De Gruyter.

Tagliamonte, Sali. 2011. *Variationist sociolinguistics: Change, observation, interpretation*. Chichester: Wiley-Blackwell.

Tagliamonte, Sali A. 2006. *Analysing sociolinguistic variation*. Cambridge: Cambridge University Press.

Vetchinnikova, Svetlana. 2015. Usage-based recycling or creative exploitation of the shared code? The case of phraseological patterning. *Journal of English as a Lingua Franca* 4(2). 223–252.

Vetchinnikova, Svetlana. 2017. On the relationship between the cognitive and the communal: A complex systems perspective. In Filppula, Markku, Juhani Klemola, Anna Mauranen & Svetlana Vetchinnikova (eds.). *Changing English: Global and local perspectives* (Topics in English Linguistics). Berlin: Mouton de Gruyter.

Weinreich, Uriel, William Labov & Marvin Herzog. 1968. Empirical foundations for a theory of language change. In Winfred P. Lehmann & Yakov Malkiel (eds.), *Directions for Historical Linguistics*, 99-188. Austin: University of Texas Press.

Wright, David. 2017. Using word n-grams to identify authors and idiolects. *International Journal of Corpus Linguistics* 22(2). 212–241.

Wulff, Stefanie, Stefan Th. Gries & Nicholas Lester. 2018. Optional that in complementation by German and Spanish learners: Where and how German and Spanish learners differ from native speakers. In Andrea Tyler, Lihong Huang & Hana Jan (eds.), *What is Applied Cognitive Linguistics: Answers from current SLA research,* 99-120. Berlin, Boston: De Gruyter Mouton.

Appendix 8.1.

Table 8.4. Syntactic categorisation used with examples

Category	Subcategory	Example	Inst. total	%
Copular:			10653	60:
(canonical)	+AdjP	*…it's a bit foolish…*[13]	3738	21
	+NP (incl. DP)	*…it's nonsense.*	5599	32
	+AdvP	*…it is enough for us.*	158	1
	+PrepP	*It's about physics.*	732	4
	+Clause	*…it's what's going on right now* *…it's because he's ignorant / It's as if*[14] *they didn't anticipate…* *…it is purely to encourage a chemical reaction.*	426	2
Progressive		*…it's already happening!*	322	2

[13] It was a deliberate strategy to choose short examples for illustration but in reality the length of the syntactic structures where *it's/is* participates varies widely. For example in copular uses, predicative complements can be very long and include various modifiers and complements, including clausal ones as in *It's as simple as measuring how much water you drink every day*.

[14] *As if* here is treated as a conjunction.

Going to V		*It's going to be a major project…*	57	0
Extraposition:			5391	31:
	inf. clause	*It's always bad to plan the future…*	2746	16
	declarative clause	*It's important that they're libertarian!*	2324	13
	interrogative clause	*…it is unclear just how this will proceed.*	315	2
	NP	*It is funny your remark…*	3	0
	gerund	*…it's nice being able to voice my opinion here.*	3	0
Passive		*…it's affected by new results.*	857	5
Cleft		*It is experimental evidence that is missing.*	377	2
Total			17,657	100

Table 8.5. Prenuclear cases and comparative constructions

Category	Subcategory	Examples	Inst. total
Prenuclear		*…who realize just how difficult it is to keep the business afloat…* *Whoever claims it is has been brainwashed…* *Ah here it is…* *…a cooling tower, no matter what kind of plant it's in, makes steam.* *…transmogrifying our banking system into the Dracula that it is today!* *So it is with all chemistry.*	157
Comparative			180
	Elliptical	*…the press is as controlled in Iran as it is in most of the world*	108
	Complement present	*…it will be faster in Firefox much like it is faster in Windows.* *The climate, as long as it is a part of science…* [15]	72
Total			337

[15] We have included cases of *as soon as, as far as* and *as long as* which often have reanalysed idiomatic meanings.